DOCUMENT RESUME

ED 379 332 TM 022 710

AUTHOR Stansfield, Charles W.; Spolsky, Bernard
TITLE Suggestions from Representatives of the International Language Testing Association for Revision of the "AERA/APA/NCME Standards for Educational and Psychological Testing."
PUB DATE Oct 94
NOTE 8p.; Testimony delivered before the Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), at the Open Conference on Revision of the "Standards" (Crystal City, VA, October 5-7, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Testing; Feedback; Language Skills; *Language Tests; *Language Usage; *Psychological Testing; Second Language Instruction; Statistical Analysis; *Test Construction; Test Reliability
IDENTIFIERS International Language Testing Association; *Performance Based Evaluation; Standard Setting; *Standards for Educational and Psychological Tests

ABSTRACT

The International Language Testing Association has some 250 members in 15 countries. Most are specialists in the testing of second language skills, with a special interest in performance assessment because of the testing of speaking and writing performance that is critical to second language skills assessment. The association believes that certain areas deserve additional attention in the next version of the "Standards for Educational and Psychological Testing." First of these is the area of standard setting. The current version of the "Standards" gives little guidance about standard-setting approaches. The experience of association members also suggests that reactions of pretest examinees can play a major role in revising and improving performance tasks. Other areas that merit further study are reliability, statistics for test analysis, and the role of language in tests. The new version of the "Standards" should address the issue of the critical role language will play in performance based tests of subject matter that rely heavily on language. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 379 332

Suggestions from Representatives of the International Language Testing Association for Revision of the *AERA/APA/NCME Standards for Educational and Psychological Testing*

by

Charles W. Stansfield, President
Second Language Testing, Inc.
10704 Mist Haven Terrace
N. Bethesda, MD 20852
Ph. (301) 231-6046
FAX (301) 231-9536

and

Bernard Spolsky, Dean
Faculty of Humanities
Bar-Ilan University
Tel Aviv, Israel
Ph. 972-2-28044
FAX 972-3-5347601

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
CHARLES W. STANSFIELD
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Testimony delivered before the Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, at the Open Conference on Revision of the *Standards*, Crystal City, Virginia, October 5-7, 1994

TM022710

BEST COPY AVAILABLE

The International Language Testing Association (ILTA) was formed in February 1992 in Vancouver, BC. The organization now has some 250 members in some 15 countries. We have a newsletter, *Language Testing Update*, and a scholarly journal, *Language Testing*. Most of our members are specialists in the testing of second language skills. Because of the special nature of our field within the educational measurement community, we may have some particular insights and concerns to contribute to the revision of the *"Standards."* However, first, we wish to point out that we are very impressed with the *Standards* document. We, like others in the measurement community, have found it to be an invaluable document as we go about our business of developing tests or conducting research on their quality. Because we are primarily concerned with the testing of the four communicative skills of listening, speaking, reading, and writing, we have considerable experience in "performance assessment." This experience derives mostly from the fact that two of these communicative skills, speaking and writing, are most often assessed by having the examinee "perform" the particular skill in question, and the performance is rated by trained raters. Given our experience in performance assessment, here are a few topics that we believe could benefit from further discussion in the next version of the *Standards*.

1. Standard setting. The current version of the *Standards* gives little guidance regarding appropriate procedures for

setting standards of passing performance on a test. There are competing approaches in terms of how it is done for multiple choice tests (Angoff, Ebel, Nedelski, etc.) and different approaches are used for performance tests. No discussion of appropriate approaches is given in the current *Standards*. Furthermore, no guidance is given as to the selection of judges or the number of judges that should be involved. Indeed, current practice typically involves a small number of judges, and yet this small number is supposed to represent everyone with a legitimate interest in the standard. Similarly, on performance assessments, current practice involves presenting the judges with a relatively small number of performances for consideration, even though these performances are supposed to represent all examinees. We believe that further guidance is needed. Such guidance would result in more valid and reliable standards on performance assessments and on high stakes tests of all types.

2. Examinee feedback. Tasks on a performance assessment are designed to elicit a response from an examinee that is an accurate indicator of his or her abilities. However, examinee performance can vary considerably across tasks. This problem is alleviated on a multiple choice test by the fact that each examinee is presented with many different items. However, one item performance tests are common. Pretesting of tasks on formal performance assessments has traditionally involved obtaining feedback on test items from raters only.

The experience of ILTA members suggests that the reactions of pretest examinees should be sought as well. Their feedback can play a major role in the revision and improvement of the tasks. If performance assessment tasks are to be valid, they must be equally fair to the vast majority of the examinees. Obtaining examinee feedback during the test development process is one way to ensure such fairness.

3. Reliability. Performance assessments, like indirect multiple-choice assessments, must be reliable. The current edition of the *Standards* gives little attention to procedures for establishing the reliability of tests involving judgements. Only one standard (2.8) deals with such matters and its call for evidence of inter-judge agreement is too simplistic. We suggest that appropriate standards for the analysis of the reliability of performance assessments be included and discussed in the next edition of the *APA Standards*.

4. Statistics for test analysis. The rift between Rasch measurement and multi-parameter IRT approaches to analyzing and equating multiple-choice tests was well established at the time the last *Standards* was published. In the interim, performance assessments have gained enormously in popularity, particularly in school districts, and new statistical procedures have been developed that are appropriate to the analysis of performance-based assessments. ILTA members have found two methodologies,

many-facet Rasch analysis and generalizability theory, to be especially useful when analyzing performance assessments involving different tasks, raters, criteria, etc. We encourage the committee that will be revising the *Standards* to mention or discuss these and other appropriate statistical methods for the analysis, scoring, and equating of performance-based tests. It is unfortunate that many large scale performance tests simply ignore the matter of equating performance-based tests or they link the score to some other related variable, such as a multiple-choice test, instead of directly equating tasks and raters to the test scale and providing scores to examinees that reflect the results of the equating.

5. The role of language in tests. Many performance tests invoke language skills to a very considerable degree. When language plays a critical role in the assessment procedure, it raises some serious questions about how we can interpret the scores. When an achievement test of a subject relies on lengthy written responses, the test is likely to be testing two abilities, achievement in the subject and language proficiency. Similarly, real world math items that are language based may also invoke language proficiency abilities that are not equally distributed in the population. The matter is a threat to validity when such performance assessments designed to assess subject matter achievement are applied to native English speakers, and the matter is a far greater threat to validity when

performance assessments are applied to nonnative speakers of English. Test developers and test score users need to consider the relative role of language in the tasks included in performance assessments.

As language testers ILTA members constantly face the flip side of this problem. We have to place our language tests within a context and familiarity with this context can easily play a role in the outcome of the language test. If the test is placed within an academic context, then performance on test tasks or items is influenced in part by knowledge of the discipline the context falls within. We have learned to edit and revise our tests very carefully to avoid invoking subject matter knowledge. However, even the efforts of professional language testers are usually less than 100% successful. We are quite aware that subject matter knowledge can influence performance on a language test that invokes that subject matter even minimally. Similarly, developers of subject matter tests must be aware of the critical role that language will play in performance based tests of subject matter that rely heavily on language. We hope the new version of the *Standards* will address this issue and extend this awareness to the entire measurement field.

ILTA and its members would be glad to take part in further discussions of these matters with the hope that they may be included in the next revision of the *Standards for Educational and Psychological Testing*.